FR-11-65

Third-Party Checking of Calibration, Scaling and Equating of the 2011 Kentucky Core Content Test

Draft until released by KDE.

Bethany H. Bynum
Arthur A. Thacker

Prepared for: **Kentucky Department of Education**
Capital Plaza Tower, 17th Floor
500 Mero Street
Frankfort, KY 40601

October 2011

HumRRO
Human Resources Research Organization
66 Canal Center Plaza, Suite 700 • Alexandria, Virginia 22314
www.humrro.org • Phone: (703) 549-3611 - Fax: (703) 519-9661

FR-11-65

Third-Party Checking of Calibration, Scaling and Equating of the 2011 Kentucky Core Content Test

Draft until released by KDE.

Bethany H. Bynum
Arthur A. Thacker

Prepared for: **Kentucky Department of Education**
Capital Plaza Tower, 17th Floor
500 Mero Street
Frankfort, KY 40601

October 2011

HumRRO
Human Resources Research Organization
66 Canal Center Plaza, Suite 700 • Alexandria, Virginia 22314
www.humrro.org • Phone: (703) 549-3611 - Fax: (703) 519-9661

EXECUTIVE SUMMARY

Two independent psychometric teams, one from Measured Progress and one from Human Resources Research Organization (HumRRO), independently estimated item parameters, scaled, and equated the 2011 Kentucky Core Content Test (KCCT). They then verified that the 2009 scoring tables could be applied to the 2011 KCCT administration for Math, Reading, Science and Social Studies. New scoring tables were produced for Writing. Procedures for scaling and equating were agreed upon by Measured Progress and HumRRO before processing began. The KCCT for Reading, Math, Science and Social studies used the same items as the 2009 administration; as a result, the psychometric process for 2011 was to verify that the 2009 scoring tables could be applied to the 2011 administration. Several procedures were conducted independently by Measured Progress and HumRRO to determine the appropriateness of applying the 2009 scoring tables to the 2011 student scores. Decisions regarding the handling of discrepancies that arose during the process of parameter estimation, scaling, equating, and the production of scoring tables were discussed between Measured Progress and HumRRO, and in all cases both groups reached consensus. Ultimately, HumRRO's recommendations to apply the 2009 scoring tables to the 2011 KCCT administration conceded with Measured Progress's recommendations. HumRRO matched Measured Progresses results for writing. Thus, HumRRO is assured that Measured Progress did not commit processing errors for Writing and that they appropriately applied the 2009 scoring tables for Math, Reading, Science and Social Studies to the 2011 test administration.

THIRD-PARTY CHECKING OF CALIBRATION, SCALING AND EQUATING OF THE 2010 KENTUCKY CORE CONTENT TEST

Table of Contents

List of Tables

THIRD-PARTY CHECKING OF CALIBRATION, SCALING AND EQUATING OF THE 2011 KENTUCKY CORE CONTENT TEST

Introduction

Every year since 1998[1], the Kentucky Core Content Test (KCCT) has been scaled and equated, and then raw-score-to-scale-score tables have been produced to be applied to students' test results. Item parameters are estimated using Item Response Theory (IRT) and then are linearly transformed (scaled) and equated (linked) with previous years' scales. The results of scaling and equating are then used to construct raw-score-to-scale-score (RSSS) tables for every KCCT test form. Cut points are also identified so that students' scores can be translated to performance categories: Novice, Apprentice, Proficient, and Distinguished (NAPD). The psychometric procedure for the 2011 KCCT was a verification process and differed slightly from prior years because the 2011 KCCT included the same items as the 2009 KCCT.

As a quality control step, the testing contractor's psychometric staff and personnel at the Human Resources Research Organization (HumRRO) simultaneously conducted an independent verification process. Researchers at both companies compared results at interim steps throughout the process, but came to independent conclusions about the ability to use the 2009 RSSS table. In this way, raw-score-to-scale-score was verified by two autonomous agencies. HumRRO has served as the third-party checker or as the primary psychometric contractor for the KCCT since the 1998–1999 academic year (Hoffman & Thacker, 1999). The results presented in this report are comparable to prior third-party investigations of KCCT psychometric processing (e.g., Sinclair, Bynum, Thacker, & Hoffman, 2008; 2007; Bynum, Sinclair, Thacker, & Hoffman, 2009).

Background

The 2007 KCCT marked the beginning a new baseline year for several reasons. First, Measured Progress replaced CTB/McGraw-Hill (CTB) as the testing contractor. Second, along with the new testing contractor came the use of different IRT software. CTB used its propriety software (PARDUX), while Measured Progress uses PARSCALE. Third, a new scale was implemented in 2007. An x0 – x80 point scale (where the "x" denotes grade) replaced the former 325 – 800 point scale. Fourth, there were test construction changes to the 2007 KCCT. Most notably, the number of open-response items was reduced and the number of multiple-choice items was increased for several grades/subjects. Fifth, the number of grades and subjects tested was expanded to meet No Child Left Behind (NCLB) guidelines[2]. Sixth, the weighting of the multiple-choice items and open-responses items changed. Prior to 2007, open-response items were weighted twice as much as multiple-choice items. However, Kentucky legislation (703 KAR 5:020) now requires that certain grades/subjects receive a 50-50 weighting, while others are weighted such that multiple-choice items account for 67% and open-response items account for 33% of the total score (Table 1 provides more detailed information on the individual

[1] The test in use before 1998 was the Kentucky Instructional Results Information System (KIRIS) test.

[2] Augmented norm-referenced tests were used to meet this requirement in 2006 to allow time for the construction of tailored criterion-referenced tests.

weighting for each test). Finally, a new standards setting took place in 2007 that resulted in new cut scores for determining performance categorizations. While none of these changes precluded the possibility of equating the 2007 KCCT to prior years, the culmination of these changes led to Measured Progress's recommendation, and the Kentucky Department of Education's (KDE) ultimate decision, to start a new trend line in 2007 and not to equate the 2007 KCCT to prior years. Consequently, the 2008 KCCT represented the first time the KCCT had been equated since the new trend line was established on the 2007 KCCT.

Table 1. Information on Individual Kentucky Core Content Tests

Grade	Subject	# MC	# OR	Weighting	OR Multiplier
3	Math	36	4	33% OR, 67% MC	1.1082089552
3	Reading	38	2	33% OR, 67% MC	2.3395522388
4	Math	32	4	50% OR, 50% MC	2.00
4	Reading	33	3	50% OR, 50% MC	2.75
4	Science	32	4	50% OR, 50% MC	2.00
4	PLVS	16	0	None	None
5	Math	32	4	50% OR, 50% MC	2.00
5	Reading	33	3	50% OR, 50% MC	2.75
5	Social Studies	32	4	50% OR, 50% MC	2.00
5	Arts & Humanities	16	1	33% OR, 67% MC	1.9701492537
5	Writing	12	2*	None	None
6	Math	32	4	50% OR, 50% MC	2.00
6	Reading	33	3	50% OR, 50% MC	2.75
7	Math	32	4	50% OR, 50% MC	2.00
7	Reading	33	3	50% OR, 50% MC	2.75
7	Science	32	4	50% OR, 50% MC	2.00
7	PLVS	16	0	None	None
8	Math	32	4	50% OR, 50% MC	2.00
8	Reading	33	3	50% OR, 50% MC	2.75
8	Social Studies	32	4	50% OR, 50% MC	2.00
8	Arts & Humanities	16	1	33% OR, 67% MC	1.9701492537
8	Writing	12	2*	None	None
10	Reading	33	3	50% OR, 50% MC	2.75
10	PLVS	16	0	None	None
11	Math	32	4	50% OR, 50% MC	2.00
11	Science	32	4	50% OR, 50% MC	2.00
11	Social Studies	32	4	50% OR, 50% MC	2.00
11	Arts & Humanities	16	1	33% OR, 67% MC	1.9701492537

Note. *For WRI05 and WRI08, there are two writing prompts. The student responds to one of the two prompts. MC = Multiple-choice items, OR = Open-response items. Grade 12 contains only prompts and is administered in the Fall.

Changes in 2010 and 2011

In 2010, the KCCT underwent several changes that impacted its psychometric processing. Kentucky, like most states, has chosen to adopt the Common Core Standards3 developed by the Council of Chief State School Officers (CCSSO) and the National Governors Association (NGA). These standards for English language arts and Mathematics are substantially different from Kentucky's Core Content, the standards on which the KCCT is based. For that reason, the 2010 and 2011 KCCT are considered transitional assessments. The assessment is based on the current Kentucky standards, but the state and the teaching field know that future assessments (beginning with the 2012 administration) will be based on the Common Core Standards.

For that reason, it was decided that additional test items based on the current Kentucky standards would not be developed or field tested. The field test positions on the KCCT were simply omitted. Also, because no new items were being developed or released, only items in the existing pool were available for use on the 2011 KCCT. In order to minimize the effort and expense of constructing and administering a new KCCT assessment, the 2011 administration was exactly the same as those administered in 2009, with the exception of Writing. Since the assessments did not change, previously constructed scoring tables could be applied with no parameter estimation, scaling, or equating steps. However, these steps were performed to verify that the assessments were functioning correctly, but ultimately the prior scoring tables were applied. Writing prompts tend to be highly memorable, as such, the Writing test in 2011 included new items. New scoring tables were computed for Writing, using similar scaling and equating procedures used in prior years.

It should also be mentioned that Kentucky's state accountability system has been placed on hold until assessments based on the Common Core Standards are available. The national accountability system under the federal NCLB legislation is still being enforced, but it is primarily focused on reading and mathematics. Kentucky also assesses science, social studies, and writing on the KCCT. These scores contribute to the "other academic indicator" part of NCLB's Adequate Yearly Progress (AYP) calculation. While these changes do not impact psychometric processing, it is possible that the reduction in the stakes for these other subjects may impact the attention they are given in schools, and could consequently impact students' test scores.

Overview of Scaling, Equating and Raw-Score-to-Scale-Score Procedures

Sample Identification and File Construction

Measured Progress does not use any exclusion rules for identifying a calibration sample. Consequently, because Measured Progress essentially uses the full student data file for IRT calibration, there is no need for HumRRO to independently construct data matrix (student data) files. Data matrix files used by both Measured Progress and HumRRO come directly from the scoring centers where students' responses are either scanned (for multiple-choice items) or hand scored (for open-response items).

Scaling Procedures

IRT scaling algorithms attempt to find item parameters that create a match between observed and theoretical response patterns as defined by the selected IRT models. Scaling produces item parameters for an achievement index called the theta scale. Item parameters for all of the 2011 test forms were estimated using PARSCALE 4.1. PARSCALE uses a three-parameter logistic model for producing parameters for multiple-choice items and a two-parameter partial credit model for producing parameters for open-response items. Item parameters from both of these models are transformed to a single theta scale with an approximate mean score of 0 and a standard deviation of 1.0.

For some items the pseudo-guessing parameter (i.e., the c-parameter) was not fully estimated during the IRT calibration. This is not an unusual occurrence, and is well-documented

in the literature (e.g., see Measured Progress Equating Report). There were also some items for which the standard error of the *b* parameter was very large (i.e., SE > .30). During the 2008 scaling and equating of the KCCT, Measured Progress and HumRRO, with the approval of KDE, agreed to the following decision rules for handling items for which the c-parameter was not fully estimated (M. Nering, email communication, June 27, 2008):

1. If estimated c = 0.0000 and the standard error of c > 0.0000, then fix c = 0.0000.
2. If standard error of b > 0.30000, then fix c = 0.0000 (the value of 0.3000 is approximate, and generally we are looking for any aberrant values).
3. Special cases. There are occasionally places where special procedures need to be implemented. These will be discussed on a case-by-case basis, and decisions for handling these problematic items will be shared in Measured Progress's watch list document.

Because the technique of fixing the c-parameter to zero typically results in stable and reasonable estimates for both the *a* and *b* parameters, this was the agreed upon default option for handling problematic items. The procedures specific to calibration are reproduced below. The full document can be found in Bynum, Thacker, Sinclair, and Hoffman, 2010.

Calibrate Data:

A. Run initial calibrations
    1. Just set and run for 100 cycles
B. Evaluate issues of non-convergence
    1. If it looks like it is converging (e.g., monotonically decreasing largest change)
        a) Increase number of cycles to allow for convergence
    2. If it is struggling with a specific item where largest change is non-monotonically decreasing
        a) If dichotomous
            (1) Fix c = 0
        b) If polytomous
            (1) Skip estimate of a parameter (SKIP = (1,0,0,0))
    3. Re-run calibration
C. Evaluate c = 0, Standard Error of c (SE(c)) > 0
    1. Set c = 0
    2. re-run calibration
D. Evaluate Standard Error of (b) (SE(b)) > 0.3 – no additional criteria
    1. Set c = 0
    2. re-run calibration
E. Re-check SE(c) > 0, c = 0; SE(b) > 0.3
    1. Set c = 0 if necessary
    2. re-run calibration if necessary
F. Evaluate model fit
    1. Examine fit plots to be sure there is no systemic problem introduced by fixing c = 0

Equating Procedures

There are two types of equating that occur for the KCCT: (a) forms equating within a given test administration year and (b) equating across test administration years using common anchor items. The first of these, forms equating, is accomplished by calibrating all of the items for a given grade/subject together. By calibrating all of the items together (i.e., across all forms), this effectively equates the various forms for a given grade/subject such that test scores on form 2 and form 3, for example, are interchangeable in terms of difficulty. In other words, a student should get about the same score regardless of which form he/she takes.

In addition to the need to equate the forms of a test within a given year, there is also the need for the current year's scores to be comparable to scores from prior years. For 2011, we equated to the 2009 scale for Reading, Math, Science and Social Studies. Kentucky uses a common-item anchor design to equate KCCT scores across years. The anchor items are "internal" in the sense that they are dispersed across forms rather than externally located in a separate anchor item form. Both multiple-choice and open-response items are designated as anchor items for equating for all grades and subjects, except for writing where only multiple-choice items are designated as anchor items (see Sinclair, Bynum, Thacker, & Hoffman, 2007).

The common-item equating process involves the application of the Stocking-Lord procedure via the Scale Transformation under Unidimensional Item Response Theory (STUIRT) models software program[3]. Upon completion of scaling, the anchor items have two sets of item parameters: (a) the item parameters from 2009, and (b) the item parameters from 2011. Since the 2011 KCCT was an exact replication of the 2009 KCCT, all of the items on the 2011 KCCT were used as anchor items. The Stocking-Lord procedure uses the prior year anchor item parameters to locate the achievement index and then searches for a transformation multiplier (M1) and an additive constant (M2) that can be combined to make the current year anchor item parameters replicate the prior year achievement index as closely as possible. This is done by attempting to match test characteristic curves (which are summations of item characteristic curves) produced by the prior year parameters with the test characteristic curves produced by transformations of current year item parameters. The transformation constants are applied to current parameters using a simple linear algebraic equation (e.g., $y = mx + b$). Since the items are the same, a comparable index should result.

Item-level reviews were conducted during the equating process. Two methods are utilized to determine if there has been anomalous change on the equating items: one based on classical test theory statistics (i.e., p-values) and the second based on IRT. The first method, the delta analysis, is the method used by Measured Progress. HumRRO independently conducts a different analysis in which the squared mean difference between quadrature points on prior year Item Characteristic Curves (ICCs) and current year ICCs are investigated. The benefit of this method is that it is not dependent on within-test-scatter—it flags items in an absolute sense, whereas the delta method flags items that are outliers relative to other items on the test. Measured Progress and HumRRO use their respective methods to independently identify items to be removed from the equating solution. Since, in 2011, we were verifying the 2009 score tables, the item-level reviews were used to determine if there were any items that had changed

[3] Downloaded from: http://www.education.uiowa.edu/casma/IRTPrograms.htm

substantially from 2009. As a result, a very conservative approach was taken for excluding items from the equating set. HumRRO investigated the squared mean difference for each item and the differences in the Item Characteristic Curves. If an item had substantially changed, then the use of the 2009 scoring tables would have been inappropriate.

Raw-Score-to-Scale-Score Procedures

Item parameters are estimated on a theta metric (mean = 0, standard deviation =1). As a result of the 2007 standards setting, three cut points were identified on the theta scale. The scaling process involves transforming the item parameters on the theta metric to the "Kentucky metric" (0 to 80 scale). The 20 and 40 cut points were fixed on the 0 to 80 scale as the cut scores separating Novice from Apprentice performance, and Apprentice from Proficient performance, respectively. These two cut scores, along with their theta values, were used to establish a linear transformation from the theta scale to the scale score metric. Consequently, the third (separating Proficient from Distinguished performance) cut point's position was defined by the line established from the other two cut points. That cut point was allowed to "float." For all grades/subjects, the first two cut scores are 20 and 40 on the scale score metric. The third cut score is not uniform across grades/subjects. The final step is to create raw-score-to-scale-score tables (i.e., "scoring tables" or "lookup tables") that can be used to convert students' number-correct score to Kentucky's reporting scale.

Verification of 2009 Scoring Tables

Two different checks were conducted to ensure the 2009 scoring tables could be applied to the 2011 data. First, the 2011 scoring tables were directly compared to the 2009 scoring tables to determine differences in raw-score-to-scale-score conversions. Second, student data was scored using the 2009 scoring tables and using the 2011 scoring tables. The results were compared to determine differences in state-level percent proficiency and state-level means for each grade and subject.

Scope of Third-Party Checking

Measured Progress and HumRRO conducted a parallel analysis to accomplish scaling, equating and the production of raw-score-to-scale-score tables for Writing. HumRRO used an independent process to verify the usability of the 2009 scoring tables. This process included scaling, equating and producing scoring tables for 2011, then comparing HumRRO's 2011 scoring tables to Measured Progress's 2009 scoring tables and HumRRO's 2011 scoring tables to Measured Progress's 2011 scoring tables. As a final verification, we compared the final 2009 score tables we received from Measured Progress in 2009 to Measured Progress's 2009 score tables that would be applied to the 2011 data. A series of specialized SAS programs were written to carry-out most of these processing steps. Below is a list of processing steps (steps where the 2009 score table verification process differed from standard procedures are noted).

Processing Steps

1. Create anchor files for equating. A SAS program produces a file for each/grade subject with prior year(s) equating item parameters to be matched with their current year item parameters.

2. Execute a SAS program to create the command files (*.psl) for PARSCALE. In order to create the command files, the SAS program reads the Item Documentation File, which contains WestEd's item identification number and the SAS data file provided by Measured Progress that contains the walk-between from WestEd's item ID to Measured Progress's item ID (i.e., "IREF" number).
3. Estimate item parameters for KCCT items using PARSCALE 4.1.
4. Check PARSCALE files for convergence and Flag items for which PARSCALE produces problematic parameters (i.e., c = 0 and/or standard error of b > .30).
5. Verify that HumRRO flags the same items as Measured Progress. Once verified, HumRRO then implements the agreed upon fixes as outlined in the Scaling procedures section.
    a) For Math, Reading, Science, and Social Studies, item flags were not verified with Measured Progress. HumRRO used the flagging procedures outlined above.
6. Rerun PARSCALE with the fixes implemented.
7. HumRRO then verifies that the PARSCALE parameter files (*.PAR) produced by Measured Progress are an identical match to the PARSCALE parameter files produced by HumRRO for Writing.
    a) This step was not conducted for Math, Reading, Science, and Social Studies.
8. Execute the SAS program that writes STUIRT command files and then execute STUIRT. STUIRT is the software used by Measured Progress to produce the Stocking-Lord constants (i.e., the multiplicative constant and additive constant) that are used to equate the current year item parameters to the prior year(s) item parameters.
9. Use the Stocking-Lord constants produced in Step 8 to equate the current year item parameters to the prior year(s) item parameters.
10. Conduct equating item analysis to determine if there are any equating items with anomalous changes that should be dropped from equating.
11. If the equating item analysis in Step 10 results in the decision to drop any equating item(s), then those items are dropped and Steps 8 and 9 are repeated before moving on to Step 10.
12. Create lookup tables for each form and each KCCT test. This places student scores on the reporting scale.
13. Confirm that the lookup tables from Step 12 are virtually identical to those derived by Measured Progress and verify the cut points used to separate student performance into NAPD categories.

Results

Scoring Table Verification

Examining the scoring tables for 2011 compared to the scoring tables for 2009 revealed that there were slight differences. Scale score differences ranged between -3 to +5. For the

majority of the grades and subjects, the differences ranged from -1 to +1 for over 90% of the scoring table, with many of the comparisons being exact matches. Differences greater than one were typically at the extreme ends of the scale where there are very few students and estimating ability is difficult. Consequently, it is not surprising that the raw-score-to-scale-score conversions fluctuate as much as five scale points at the ends of the scales (e.g., 300 to 305). That difference makes very little practical difference in estimating school-level annual yearly progress.

The difference in the state means using the 2011 scoring tables and the 2009 scoring tables ranged from -1.39 to 1.18, with a mean difference across grades and subjects of 0.46 and a median of 0.44 (see Table 2). The percent difference of the state means ranged from 0% to 0.68%, with a mean of 0.23% and median of 0.25%. Given the slight difference in the scoring tables, the differences in state means are not unexpected. However, the differences are very small. Similarly, differences in state-level percentage of student proficiency ranged from -1.40% to 1.73%, with a mean of 0.93% and median of 0.81%. The differences were not systematic, such that for some grades, the 2011 scoring table yielded slightly higher scores and in other grades, the 2009 scoring tables yielded slightly higher scores. Overall, the differences between applying the 2011 scoring tables and applying the 2009 scoring tables are within an acceptable margin of error.

Table 2. Summary of the Verification of the 2009 Scoring Tables

Subject	Grade	% proficient			State Means			
		2011	2009	diff	2011	2009	Diff	% Diff[1]
Math	3	77.65	78.15	0.50	355.60	356.06	0.46	0.21%
	4	73.98	75.15	1.17	453.32	453.89	0.57	0.26%
	5	67.05	66.24	-0.81	549.25	548.93	-0.32	0.16%
	6	69.3	70.1	0.80	648.94	649.43	0.49	0.25%
	7	66.37	65.97	-0.40	748.04	747.47	-0.57	0.30%
	8	60.77	59.98	-0.79	844.57	844.46	-0.11	0.06%
	10	65.24	66.16	0.92	1045.37	1045.81	0.44	0.24%
Reading	3	79.26	80.88	1.62	354.91	355.84	0.93	0.42%
	4	75.28	73.88	-1.40	452.09	450.70	-1.39	0.68%
	5	72.91	74.42	1.51	550.75	551.74	0.99	0.48%
	6	70.26	71.99	1.73	649.50	650.67	1.18	0.59%
	7	66.5	67.16	0.66	747.29	747.50	0.21	0.11%
	8	71.43	71.84	0.41	848.73	848.77	0.04	0.02%
	11	46.62	46	-0.62	1136.75	1136.75	0.00	0.00%
Science	4	70.27	70.85	0.58	449.80	449.73	-0.07	0.04%
	7	63.5	64.5	1.00	745.72	745.75	0.03	0.01%
	11	40.82	41.35	0.53	1135.61	1135.72	0.11	0.08%
Social Studies	5	59.32	59.87	0.55	543.81	544.31	0.50	0.29%
	8	58.41	60.07	1.66	843.59	844.04	0.45	0.25%
	11	42.28	41.24	-1.04	1136.07	1135.71	-0.36	0.25%

Note.
[1]Percent difference is based on the difference between the 2011 and 2009 mean score minus the base year (i.e., for grade 3, the percent difference was based on 55.60 for 2011 and 56.06 for 2009).

Writing

Table 3 summarizes the results of the third-party checking for Writing. The 3rd column indicates whether any actions were taken to handle items for which PARSCALE produced problematic parameters. Those actions were implemented and thereby produced unequated parameter files that were identical to the unequated parameter files produced by Measured Progress (see column 4). The next step was to run STUIRT and produce transformation constants (M1 and M2) for equating the 2011 KCCT to the prior test year. The equated parameters produced by HumRRO were an exact match to the equated parameters produced by Measured Progress (see column 5). Finally, the look-up tables (i.e., raw-score-to-scale-score tables) were produced along with NAPD classifications. As seen in columns 6 and 7, HumRRO's NAPD classifications and look-up tables were an exact match with Measured Progress's NAPD classifications and look-up tables for all grades/subjects. Because HumRRO and Measured Progress reached exact agreement on calibration results and lookup tables, while using different methods to produce those results, HumRRO is assured that the results are accurate and that Measured Progress did not commit processing errors.

Table 3. Summary of 2011 Results – Writing

SUB	GR	Check Item list (MATCH MP?)	Check Item Parameters (Item with FLAGS?)	Why Flag??	Match Fixes	Match MPs PAR files	EQ ITEM DROPS	Match RSSS
WR	5	YES	32394	c = 0	YES[1]	YES	NONE	YES
			30508	se_b > 3	NO[2]			
			32319	se_b > 3	NO[2]			
WR	8	YES	31405	se_b > 3	NO[2]	YES	NONE	YES

Note.
[1]Measured Progress and HumRRO agreed to fix item.
[2]After discussing the discrepancies, Measured Progress and HumRRO agreed not to fix these items.

Documentation[4]

To document the steps involved in scaling and equating of the 2011 KCCT, HumRRO saved all electronic files used in data preparation, such as SAS files, and all files produced during PARSCALE calibration. The core 2011 files have been submitted to KDE in electronic format.

All KCCT electronic files submitted to KDE are named according to the code listed below (G = grade level, S = subject).

A. Data Matrix File (SSSGGMatrix.TXT). This file contains the student score data. It is coded such that a "1" indicates a correct answer for a multiple-choice question and actual score levels are recorded for student responses to open-response questions. For all grades/subjects, except writing, the score levels for open-response items are 0 – 4. For writing, because there were always two raters, the score levels for open-response

[4] Note that due to the manner in which open-response items are handled for WRI05 and WRI08, these SAS programs had to be altered to handle writing. Consequently, there are separate, but analogous, programs to handle WRI05 and WRI08.

items are either 0 – 8 or 2 – 8 (one writing dimension was scored 1 – 4 rather than 0 – 4 by two raters and each prompt is scored for 3 dimensions). The data in the data matrix file is ordered such that common multiple-choice items are followed by matrix multiple-choice items in form 1 by item number, in form 2 by item number, etc., which are followed by common open-response items, which are finally followed by matrix open-response items in form 1 by item number, in form 2 by item number, etc.

B. SAS program, "0 Create Anchor Files.SAS." This program produces a file for each/grade subject with prior year(s) equating item parameters to be matched with their current item parameters.

C. SAS program, "1a Itemdoc & Parscale file.SAS." This SAS program reads the Item Documentation File and the item ID walk-between file and creates PARSCALE command files. This SAS program automates the creation of PARSCALE command files for all grades/subjects except writing.

D. PARSCALE Command File (10GGSSS.PSL). These file contains the number of items for which parameters are being estimated, the maximum number of iterations for convergence, starting values for the c-parameter, the response model to be used (graded or partial), and the response function metric to be used (logistic or normal). It also contains information allowing the program to distinguish between open-response and multiple-choice items, and the number of score levels for open-response items.

E. PARSCALE Output File Phase 0 (09GGSSS.PHO). This file evaluates the PARSCALE command file against the data matrix file.

F. PARSCALE Output File Phase 1 (09GGSSS.PH1). This file starts the processing of the data matrix file. It displays the frequencies for correct and incorrect responses by item. The file also outputs point biserials.

G. PARSCALE Output File Phase 2 (09GGSSS.PH2). This file displays the history of the parameter estimation phase. It lists a systematic iteration of data, by item, during each stage of parameter estimation. This file is helpful in determining for which items PARSCALE is having problems estimating parameters.

H. PARSCALE Parameter File (SSS09GG.PAR). This file contains parameter estimates for all items designated in the *.PSL file. It is used for later data manipulation.

I. PARSCALE Fit File (SSS09GG.FIT). This file contains fit statistics for all items.

J. Output from SAS program, "2a Check IRT parameters" This SAS program reads in the PAR file produced by PARSCALE and flags items that are not discriminating ($a < 0$), that are "easy" ($b < -2.0$), that are "hard" ($b > 2.0$), that are easy to guess ($c > .40$), and, most importantly, items with zero c-parameters and items for which the standard error of $b > .30$. A secondary purpose of this program is to split the PAR files produced by PARSCALE into individual forms for each grade/subject; this is the format needed to produce raw-score-to-scale-score tables. Finally, this file flags any items for which HumRRO and Measured Progress do not produce identical unequated parameters.

J.1 For Math, Reading, Science, and Social Studies, interim checks with Measure Progress's results were not conducted. The documentation for these subjects for 2011 only continues the Checks of IRT parameters.

K. Output from SAS program, "3a equate across years with STUIRT.SAS." This program uses the anchor files that were created by the program, "0 Create Anchor Files.SAS" and matches those items with their current year item parameters. It then writes and executes STUIRT command files. Then, it reads the STUIRT output file and searches for the Stocking-Lord constants. It then uses the Stocking-Lord constants to create the current year equated item parameters. Next, this program performs delta analyses. It computes delta based on: (a) prior year anchor item p-values from Measured Progress's equate files and (b) current year anchor item p-values computed from frequency counts in the PARSCALE phase 2 output file. This program then compares HumRRO's computed deltas to the deltas reported by Measured Progress. Finally, this program plots ICCs from the current year and the anchor year and computes the mean squared difference between the curves so as to identify items with anomalous change.

L. Output from SAS program, "4a Compare MP & HumRRO equating.SAS." This program compares HumRRO's equated parameters with Measured Progress's equated parameters

L.1 For Math, Reading, Science, and Social Studies, interim checks with Measure Progress's results were not conducted. This step was not conducted for these grades.

M. Output from SAS program, "5a raw-score-to-scale-score 80 point version.SAS." This SAS program performs the raw-score-to-scale-score transformations using HumRRO's boundary method. Lower boundaries are created for each raw score and, using the expected score, the theta value is classified into a raw score boundary. Lookup tables for each KCCT form are then produced. This program outputs the max score for each grade subject and the weight multiplier for open-response items.

N. Raw-Score-to-Scale-Score Tables/Lookup Tables (SSS09GGrssshrX.SAS). A raw-score-to-scale-score table is produced for each KCCT form (the "X" at the end of the file name is a placeholder to denote form).

O. Output from SAS program, "6a Compare MPresults with HumRRO results—RSSS.SAS." This SAS program compares the lookup tables produced by HumRRO to the lookup tables produced by Measured Progress and outputs any discrepancies in scale scores and NAPD classifications.

O.1 Three different comparisons were made, including: (a) comparing HumRRO's 2011 scoring tables to Measured Progress's 2009 scoring tables, (b) comparing HumRRO's 2011 scoring tables to Measured Progress's 2011 scoring tables and (c) comparing the final 2009 score tables we received from Measured Progress in 2009 to Measured Progress's 2009 score tables that would be applied to the 2011 data.

Conclusion

Measured Progress and HumRRO independently scaled and equated the KCCT and verified the use of 2009 scoring tables for the 2011 administration. For writing, new scoring tables were produced for the 2011 administration. From these lookup tables, cut points were identified that can be used to assign student performance classifications NAPD and convert to school accountability indexes. Decisions regarding the handling of discrepancies that arose during the process of scaling and equating were discussed between Measured Progress and HumRRO and in all cases both groups reached consensus. In the end, for writing, results calculated by HumRRO were identical to those calculated by Measured Progress. For Math, Reading, Science, and Social Studies, both HumRRO and Measured Progress independently concluded applying the 2009 scoring tables to the 2011 test administration was appropriate. Further, HumRRO verified the 2009 score tables applied to the 2011 administration against those produced in 2009. Given that, our recommendations to apply the 2009 scoring tables to the 2011 KCCT administration conceded with Measured Progress's recommendations and our results exactly matched Measured Progresses results for writing, we are assured that Measured Progress did not commit processing errors for writing and that they appropriately applied the 2009 scoring tables for Math, Reading, Science and Social Studies to the 2011 test administration.

References

Bynum, B.H., Sinclair, A.L, Thacker, A.A., & Hoffman, R.G. (2009). *Third-party checking of 2009 scaling and equating of the Kentucky core content test* (FR-09-79). Alexandria, VA: Human Resources Research Organization.

Bynum, B.H., Thacker, A.A., Sinclair, A.L, & Hoffman, R.G. (2010). *Third-party checking of 2010 scaling and equating of the Kentucky core content test* (FR-10-62). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G. & Thacker, A. A. (1999). Third-party checking of 1999 scaling and linking for the Kentucky Core Content Test. (HumRRO Report SP-WATSD-99-44). Alexandria, VA: Human Resources Research Organization.

Sinclair, A.L, Bynum, B.H., Thacker, A.A., & Hoffman, R.G. (2008). Third-party checking of 2008 scaling and equating of the Kentucky core content test (FR-08-86). Alexandria, VA: Human Resources Research Organization.

Sinclair, A.L, Bynum, B.H., Thacker, A.A., & Hoffman, R.G. (2007). Third-party checking of 2007 scaling and equating of the Kentucky core content test (FR-07-54). Alexandria, VA: Human Resources Research Organization.